Vector-Valued Support Vector Regression

Mark Brudnak, *Member, IEEE*

Abstract—A vector-valued extension of the support vector regression problem is presented here. The vector-valued variant is developed by extending the notions of the estimator, loss function and regularization functional from the scalar-valued case. A particular emphasis is placed on the class of loss functions chosen which apply the ε-insensitive loss function to the p-norm of the error. The primal and dual optimization problems are derived and the KKT conditions are developed. The general case for the p-norm is specialized for the 1-, 2- and ∞-norms. It is shown that the vector-valued variant is a true extension of the scalar-valued case. It is then shown that the vector-valued approach results in sparse representations in terms of support vectors as compared to aggregated scalar-valued learning.

I. INTRODUCTION

Multi-task learning is concerned with mappings of the form $f : \mathbb{R}^n \mapsto \mathcal{Y}^m$ where $\mathcal{Y} \triangleq \{0, 1\}$ for classification and $\mathcal{Y} \triangleq \mathbb{R}$ for regression. The problem of multi-task learning is approachable as an aggregation of independent single-task learning problems $f_i : \mathbb{R}^n \mapsto \mathcal{Y}$. Certainly, there is no loss of generality with this approach, however, when the number of tasks, m, is large, the aggregated approach has some disadvantages. Regardless of the single-task method used, the aggregated method requires m optimizations, which for the support vector machine (SVM), potentially requires m sets of redundant kernel computations. Also, for SVM, it is likely that there will only be coincidental commonality between the support vectors associated with the different tasks. The impact of this first disadvantage is incurred during training, however, the cost of the second is incurred during use. Both of these costs may be negligible where kernel computations are inexpensive, however as kernel computations become more expensive (i.e. large n) these costs may become significant. Another motivation according to Micchelli and Pontil [1] for multi-task learning is mutual dependence among the tasks, which the aggregated approach ignores. Such dependence occurs when $y = f(x)$ is an embedding (which is certain when $n < m$). In such a case, the outputs y lie on a subspace embedded in $\mathbb{R}^m$. For regression such embeddings occur with the use of the unit quaternions to represent rotations in 3-space, which are locally parameterized by three orthogonal coordinates, but embedded in $\mathbb{R}^4$. They also occur in the configuration space of mechanisms with closed kinematic chains for which a global parameterization is not available. These cases as well as others motivate the development of multi-task learning methods and in particular the multi-task SVM for which the

Mark Brudnak is a researcher with the U.S. Army RDECOM-TARDEC, 6501 E. 11 Mile Road, Warren, MI 48397-5000, USA (phone: 586-574-7355; email: brudnakm@tacom.army.mil).

regression problem will be addressed here. Such regression problems are described as *vector-valued* in the sequel.

Recent work on vector-valued support vector regression (VV-SVR) is as follows. Vazquez and Walter [2] use a separately trained Matérn class kernel for each task and then use the single-task SVM for training. Micchelli and Pontil [1], [3] give a theoretical treatment of reproducing kernel Hilbert spaces (RKHS) in the range-space of the estimator. Their result is an extension of the traditional scalar-valued kernel function to an *operator-valued* kernel. Ben-David and Schuller [4] develop conditions under which learning multiple tasks is provably beneficial. Evgeniou and Pontil [5] consider the learning of an average task simultaneously with small deviations for each task. Evgeniou, Micchelli and Pontil [6] extend their earlier results by developing indexed kernels with coupled regularization functionals.

In contrast to this previous work, this paper emphasizes the choice of loss function in the vector-valued regression problem. Prior work on the loss function by Pérez-Cruz et al [7] used the squared Euclidian norm of the error with a hyper-spherical insensitive zone. Also, Sánchez-Fernández, et al [8] used a shifted squared Euclidian norm for a differentiable loss function. These two approaches do not reduce to the traditional SVR loss function in one-dimensional cases. The VV-SVR proposed here generalizes the ε-insensitive loss function of the scalar-valued case. It follows the traditional scalar-valued SVM development [9]. The problem is first setup by defining a regularized risk functional which extends the scalar-valued case. This problem is then cast into primal, Lagrangian and then dual forms. We then develop the Karush-Kuhn-Tucker (KKT) conditions to relate the dual variables to the primal variables which are used to find the bias. The general case is then specialized for the common norms and specific approaches to determination of the bias are developed. The method is demonstrated and shown to be sparse in support vectors. The paper concludes with a comparison of the vector-valued case to the scalar-valued case and some observations.

II. SCALAR-VALUED SUPPORT VECTOR REGRESSION

In the scalar-valued support vector regression (SV-SVR) problem one seeks to model a causal relationship $f : \mathbb{R}^n \mapsto \mathbb{R}$ between inputs x and an output y from a finite set of observations $\{(x_i, y_i)\}_1^\ell$. Generally such a regression problem takes the form of

$$\text{Min}_{\boldsymbol{\pi}} : \quad R_{reg} = \mathcal{P}(\boldsymbol{\pi}) + C \sum_{i=1}^{\ell} L\left(y_i, \hat{y}\left(x_i, \boldsymbol{\pi}\right)\right) \quad (1)$$

in which we wish to minimize both the summed loss and a regularization functional simultaneously. For the SV-SVR

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
14 APR 2006	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Vector-Valued Support Vector Regression	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Brudnak, Mark	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
USA TACOM 6501 E 11 Mile Road Warren, MI 48397-5008	**15692**

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	TACOM TARDEC
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at IEEE International Joint Conference on Neural Networks 2006

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	**SAR**	**8**	
unclassified	**unclassified**	**unclassified**			

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

problem the estimator $\hat{y}(\cdot,\cdot)$ is chosen as

$$\hat{y}(\mathbf{x}, \{\mathbf{w}, b\}) = \langle \mathbf{w}, \ \phi(\mathbf{x}) \rangle + b \qquad (2)$$

where $\phi : \mathbb{R}^n \mapsto \mathbb{R}^\nu$ (or $\phi : \mathbb{R}^n \mapsto L^2$) is a nonlinear mapping to a high-dimensional *feature space* and clearly $\pi \triangleq \{\mathbf{w}, b\}$ where $\mathbf{w}$ is the *weight vector* and b is the *bias*. We desire that the SV-SVR perform well on our set of observations, so we choose $\mathbf{w}$ and b so that the summed loss $\sum_i L(y_i, \hat{y}_i)$ is minimized where the loss function $L(\cdot, \cdot)$ is typically the ε-insensitive loss function $L(y, \hat{y}) \equiv L(e) = |e|_\varepsilon \triangleq \max(0, |e| - \varepsilon)$ where $e \triangleq y - \hat{y}$. Since $\nu \gg \ell$, the minimization of the summed loss alone is ill-posed and therefore traditional SVR introduces the regularizational functional $\mathcal{P}(\mathbf{w}) \triangleq \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle$ to stabilize the solution. With these choices, the SV-SVR optimization problem in (1) becomes

$$\begin{aligned}\underset{\{\mathbf{w}, b\}}{\text{Min}} \ : \ R_{reg} = &\frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle \\ &+ C \sum_{i=1}^{\ell} |y_i - \langle \mathbf{w}, \ \phi(\mathbf{x}_i) \rangle - b|_\varepsilon .\end{aligned} \qquad (3)$$

Now because this problem is cast in a large space of parameters $\{\mathbf{w}, b\}$ and because it is non-smooth, it is transformed into the dual problem

$$\begin{aligned}\underset{\{\beta_i\}_1^\ell}{\text{Max}} \ : \quad D = &-\frac{1}{2} \sum_{i,j=1}^{\ell} \beta_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{\ell} y_i \beta_i \\ &- \varepsilon \sum_{i=1}^{\ell} |\beta_i|\end{aligned} \qquad (4)$$

$$\text{S.T.:} \quad \sum_{i=1}^{\ell} \beta_i = 0, \quad |\beta_i| \le C .$$

where $k(\mathbf{x}_i, \mathbf{x}_j) \triangleq \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ is the kernel function. Given the fact that $\mathbf{w} = \sum_{i=1}^{\ell} \beta_i \phi(\mathbf{x}_i)$, the final form of the estimator then becomes

$$\hat{y}(\mathbf{x}) = \sum_{i \in I_{SV}} \beta_i k(\mathbf{x}_i, \mathbf{x}) + b \qquad (5)$$

where $I_{SV} \triangleq \{i : \beta_i \ne 0\}$ denotes the set of indices of the support vectors $\mathbf{x}_i$ and b is determined using the KKT conditions, which in the SV-SVR case may be briefly stated as

$$\begin{aligned}\beta_i = 0 &\implies |e_i| < \varepsilon, \\ 0 < |\beta_i| < C &\implies |e_i| = \varepsilon, \\ |\beta_i| = C &\implies |e_i| > \varepsilon, \\ \beta_i \ne 0 &\implies \beta_i e_i > 0.\end{aligned}$$

This SV-SVR estimator has several desirable properties: (1) it generalizes well, (2) it is based on linear math, (3) the optimization problem is convex and (4) the dual problem is quadratic. In particular, the SVR's generalization ability is attributable to the sparsisity and boundedness of the dual problem (4) solution. Sparsisity is attributable to the ε-insensitive zone (i.e. $-\varepsilon < e_i < \varepsilon$) of the loss

function since the $\varepsilon \sum_{i=1}^{\ell} |\beta_i|$ term creates a cusp in the objective function (4), "trapping" solutions at $\beta_i = 0$. Both the ε-insensitive zone and the cusp vanish if $\varepsilon = 0$. The boundedness of each β_i is attributable to the linear part of the loss function (i.e. the bound on β_i is related to $C \frac{\partial L}{\partial e}$, see Smola and Schölkopf [10, pg. 13]). Since both the sparsisity and the boundedness of the estimator limit the number of free parameters of the estimator, in a practical sense the generalization properties of SV-SVR may be attributable to the form of the loss function.

III. VECTOR-VALUED SUPPORT VECTOR REGRESSION

We will now extend those concepts just discussed for the SV-SVR case to the vector-valued case. For each choice made we will maintain the concepts from the scalar-valued case to assure that the VV-SVR is a true generalization of the SV-SVR.

A. Problem Setup

In the vector-valued case, the process to be estimated is of the form $\mathbf{f} : \mathbb{R}^n \mapsto \mathbb{R}^m$ which maps inputs $\mathbf{x} \in \mathbb{R}^n$ to vector-valued outputs $\mathbf{y} \in \mathbb{R}^m$. From a finite set observations $\{(\mathbf{x}_i, \mathbf{y}_i)\}_1^\ell$, our goal is to find a function $\hat{\mathbf{y}}(\mathbf{x}, \pi)$ which will be trained over its free parameters π as

$$\underset{\pi}{\text{Min}} \ : \quad R_{reg} = \mathcal{P}(\pi) + C \sum_{i=1}^{\ell} L(\mathbf{y}_i, \hat{\mathbf{y}}(\mathbf{x}_i, \pi)). \qquad (6)$$

The structure of $\mathcal{P}(\cdot)$, $\hat{\mathbf{y}}(\cdot, \pi)$ and $L(\cdot, \cdot)$ are now chosen to extend the same concepts from the SV-SVR case. For VV-SVR, the family of functions $\hat{\mathbf{y}}(\cdot, \cdot)$ will take the linear form

$$\hat{\mathbf{y}}(\mathbf{x}; \{\mathbf{W}, \mathbf{b}\}) \triangleq \mathbf{W}\phi(\mathbf{x}) + \mathbf{b} \qquad (7)$$

which generalizes (2) where the free parameters $\pi = \{\mathbf{W}, \mathbf{b}\}$ consist of the *weights* $\mathbf{W} \in \mathbb{R}^{m \times \nu}$ and the *bias* $\mathbf{b} \in \mathbb{R}^m$. Similar to the SV-SVR case, a quadratic regularization functional is chosen as $\mathcal{P}(\mathbf{W}) \triangleq \frac{1}{2} \text{Tr}(\mathbf{W}\mathbf{W}^T)$. Finally, the loss function must be extended such that $L(\cdot, \cdot) : \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathbb{R}_+$. To construct this loss function in the spirit of the SV-SVR, it is natural to maintain the concept of the ε-insensitive loss function. Such an extension is obtained by applying $|\cdot|_\varepsilon$ to a norm of the error $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ which yields

$$L(\mathbf{e}) \triangleq \Big| \|\mathbf{e}\|_p \Big|_\varepsilon \qquad (8)$$

where $\mathbf{e} = [e_1 \ \cdots \ e_m]^T$ and $\|\mathbf{e}\|_p$ is the p-norm defined by $(\sum_{i=1}^m |e_i|^p)^{\frac{1}{p}}$ for $1 \le p < \infty$ and $\max_i |e_i|$ for $p \sim \infty$. This loss function has an insensitive zone for $\Big| \|\mathbf{e}\|_p \Big|_\varepsilon \le \varepsilon$, is a true generalization of $|\cdot|_\varepsilon$, and has a "linear" behavior for $\Big| \|\mathbf{e}\|_p \Big|_\varepsilon > \varepsilon$. Combining these choices for $\hat{\mathbf{y}}(\cdot, \cdot)$, π, $\mathcal{P}(\cdot)$ and $L(\cdot)$, (6) becomes

$$\begin{aligned}\underset{\{\mathbf{W}, \mathbf{b}\}}{\text{Min}} \ : \quad R_{reg} = &\frac{1}{2} \text{Tr}(\mathbf{W}\mathbf{W}^T) \\ &+ C \sum_{i=1}^{\ell} \Big| \|\mathbf{y}_i - \mathbf{W}\phi(\mathbf{x}_i) - \mathbf{b}\|_p \Big|_\varepsilon\end{aligned} \qquad (9)$$

which we note is a generalization of (3). In the next section we will convert (9) into a suitable form for practical optimization by developing the dual problem in the space of Lagrange multipliers.

B. Dual Problem Development

Observe that (9) is non-smooth and may be infinite dimensional; however, it may be simplified by deriving the dual problem. First, the objective function may be smoothed by introducing the slack variables ξ_i, δ_i and δ_i^* resulting in the primal optimization problem

$$
\begin{aligned}
\underset{\mathbf{W},\,\mathbf{b},\,\{\xi_i,\,\delta_i,\,\delta_i^*\}_1^\ell}{\text{Min:}} \quad & P \triangleq \frac{1}{2}\operatorname{Tr}(\mathbf{W}\mathbf{W}^T) + C\sum_{i=1}^{\ell}\xi_i \\
\text{S.T.:} \quad & \|\delta_i + \delta_i^*\|_p - \xi_i - \varepsilon \le 0,\ \xi_i \ge 0, \\
& \mathbf{y}_i - \mathbf{W}\phi(\mathbf{x}_i) - \mathbf{b} - \delta_i \le 0, \\
& -\mathbf{y}_i + \mathbf{W}\phi(\mathbf{x}_i) + \mathbf{b} - \delta_i^* \le 0, \\
& \delta_i \ge 0,\ \delta_i^* \ge 0,\ \forall\, i = 1,\ldots,\ell
\end{aligned}
\tag{10}
$$

where the inequalities are taken element-wise and $\|\cdot\|_p$ is defined without the absolute value since δ_i, $\delta_i^* \ge 0$. Now to reduce the dimension of the problem, we will cast the primal problem P into Lagrange form by introducing the multipliers, α_i, η_i, γ_i, γ_i^*, θ_i, θ_i^* as

$$
\begin{aligned}
L \triangleq\ & \frac{1}{2}\operatorname{Tr}(\mathbf{W}\mathbf{W}^T) + C\sum_{i=1}^{\ell}\xi_i \\
& + \underbrace{\sum_{i=1}^{\ell}\alpha_i\left(\|\delta_i + \delta_i^*\|_p - \xi_i - \varepsilon\right)}_{\text{III}} \\
& - \underbrace{\sum_{i=1}^{\ell}\eta_i\xi_i}_{\text{IV}} + \underbrace{\sum_{i=1}^{\ell}\gamma_i^T\left(\mathbf{y}_i - \mathbf{W}\phi(\mathbf{x}_i) - \mathbf{b} - \delta_i\right)}_{\text{V}} \\
& + \underbrace{\sum_{i=1}^{\ell}\gamma_i^{*T}\left(-\mathbf{y}_i + \mathbf{W}\phi(\mathbf{x}_i) + \mathbf{b} - \delta_i^*\right)}_{\text{VI}} \\
& - \underbrace{\sum_{i=1}^{\ell}\theta_i^T\delta_i}_{\text{VII}} - \underbrace{\sum_{i=1}^{\ell}\theta_i^{*T}\delta_i^*}_{\text{VIII}} \\
\text{S.T.:}\ & \alpha_i \ge 0,\ \eta_i \ge 0,\ \gamma_i^{(*)} \ge 0,\ \theta_i^{(*)} \ge 0
\end{aligned}
\tag{11}
$$

which is to be minimized over $\mathbf{W}$, $\mathbf{b}$, ξ_i and $\delta_i^{(*)}$ and to be maximized over α_i, η_i, $\gamma_i^{(*)}$ and $\theta_i^{(*)}$ for $i = 1,\ldots,\ell$. Now since the primal variables are no longer constrained and all constraints are imposed on the dual variables, we minimize (11) with respect to the primal variables. Minimizing with

respect to $\mathbf{W}$, $\mathbf{b}$, $\delta_i^{(*)}$ and ξ_i yields

$$
\mathbf{W} = \sum_{i=1}^{\ell}\Gamma_i\phi^T(\mathbf{x}_i)
\tag{12}
$$

$$
\sum_{i=1}^{\ell}\Gamma_i = \mathbf{0}
\tag{13}
$$

$$
\theta_i^{(*)} = \alpha_i\frac{\partial}{\partial\delta_i^{(*)}}\left(\|\delta_i + \delta_i^*\|_p\right) - \gamma_i^{(*)}
\tag{14}
$$

$$
\eta_i = C - \alpha_i
\tag{15}
$$

respectively where $\Gamma_i \triangleq \gamma_i - \gamma_i^*$. These relationships along with the constraints $\eta_i \ge 0$ and $\theta_i^{(*)} \ge 0$ imply that

$$
0 \le \alpha_i \le C
\tag{16}
$$

$$
0 \le \gamma_i^{(*)} \le \alpha_i\frac{\partial}{\partial\delta_i^{(*)}}\left(\|\delta_i + \delta_i^*\|_p\right).
\tag{17}
$$

Observe that the results expressed in (12) and (13) are similar to those for the SV-SVR. By combining (12) with $\hat{\mathbf{y}}$ in (7) the expression

$$
\hat{\mathbf{y}}(\mathbf{x}) = \sum_{i=1}^{\ell}\Gamma_i\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + \mathbf{b} = \sum_{i=1}^{\ell}\Gamma_i k(\mathbf{x}_i,\mathbf{x}) + \mathbf{b}
$$

of the VV-SVR estimator is obtained which may be compared to (5). Now the conditions expressed in (17) reflect a complicated coupling between the primal and dual variables which may be simplified with the following lemma.

Lemma 1: Let $\mathcal{V} = (\mathbb{R}^n, \|\cdot\|_p)$ define a normed vector space and let $\chi \in \mathcal{V}$. Also, let $\|\cdot\|_q$ be the dual or conjugate norm of $\|\cdot\|_p$ (that is $\frac{1}{p} + \frac{1}{q} = 1$). Then

$$
\left\|\frac{d}{d\chi}\left(\|\chi\|_p\right)\right\|_q \equiv 1
$$

for $1 \le p \le \infty$.

Proof: We will examine three cases. We begin with the smooth case and then address the non-smooth cases where $p \in \{1, \infty\}$.

Case 1: Consider $1 < p < \infty$ for which $\|\cdot\|_p$ is smooth everywhere except at $\mathbf{0}$. In this case for $\chi = \begin{bmatrix}\chi_1 & \cdots & \chi_n\end{bmatrix}^T \in \mathbb{R}^n$ we have $\|\chi\|_p \triangleq (\sum_{i=1}^{n}|\chi_i|^p)^{\frac{1}{p}}$ then $\frac{d}{d\chi}\left(\|\chi\|_p\right) = \frac{1}{\|\chi\|_p^{p-1}}\begin{bmatrix}\pm|\chi_1|^{p-1} & \cdots & \pm|\chi_n|^{p-1}\end{bmatrix}^T$. Now since $q = \frac{p}{p-1}$, direct manipulation yields

$$
\begin{aligned}
\left\|\frac{d}{d\chi}\left(\|\chi\|_p\right)\right\|_{\frac{p}{p-1}} &= \left(\sum_{i=1}^{n}\left(\frac{|\chi_i|^{p-1}}{\|\chi\|_p^{p-1}}\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} \\
&= \left(\frac{1}{\|\chi\|_p^p}\sum_{i=1}^{n}\left(|\chi_i|^p\right)\right)^{\frac{p-1}{p}} \\
&= 1.
\end{aligned}
$$

In the following two cases the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are non-smooth. However, it will be shown that norm may be modeled as a collection of smooth subspaces which cover $\mathcal{V}$.

3

The gradient may then be calculated in each of these subspaces to yield all of the possible directions of *maximum ascent* at a non-smooth point. In this sense the gradient is multi-valued at non-smooth points, however, it is shown that the dual norm of each of these possible values is always 1.

Case 2: Next, consider $p = 1$. In the smooth regions $\frac{d}{d\chi} \|\chi\|_1 = [\pm 1 \quad \cdots \quad \pm 1]$ which clearly has an ∞-norm of 1. Non-smooth regions exist for any $\chi_i = 0$ (called the *coordinate subspaces* here). Let χ_{ns} denote a candidate point in one of these coordinate subspaces. We define the smooth coordinate vector $\mathcal{C}(\chi_{ns}) \triangleq [\mathcal{C}_1 \quad \cdots \quad \mathcal{C}_m]^T$ as

$$\mathcal{C}_i(\chi_{ns}) = \begin{cases} 0, & \text{if } \chi_i = 0 \\ \text{sign}(\chi_i), & \text{if } \chi_i \neq 0 \end{cases},$$

the *zero coordinate vector* as $\mathcal{Z}(\chi_{ns}) = 1 - |\mathcal{C}|$ and the *ternary permutation vector* as $\mathcal{T}(\chi_{ns}) = [\mathcal{T}_1 \quad \cdots \quad \mathcal{T}_m]^T$ where $\mathcal{T}_i \in \{-1, 0, 1\}$, $\forall \ i = 1, \ldots, m$. In the neighborhood of a non-smooth point χ_{ns}, the norm may be modeled as the linear functional given by $\|\chi\|_1 = (\mathcal{C}(\chi_{ns}) + \mathcal{Z}(\chi_{ns}) \circ \mathcal{T}(\chi_{ns}))^T \chi$ where $\circ$ denotes an element-wise product. It is clear that all of the directions of maximum ascent (defined by the permutations of $\mathcal{Z}(\cdot) \circ \mathcal{T}(\cdot)$) from such a non-smooth point are given by

$$\frac{d}{d\chi} (\|\chi\|_1) \Big|_{\chi = \chi_{ns}} \in \{\mathcal{C}(\chi_{ns}) + \mathcal{Z}(\chi_{ns}) \circ \mathcal{T}(\chi_{ns})\} \backslash \{0\}.$$

all of which have an ∞-norm of 1, even for $\chi_{ns} = 0$.

Case 3: Finally consider $p \sim \infty$. In smooth regions $\frac{d}{d\chi} \|\chi\|_\infty = [\cdots \quad 0 \quad \pm 1 \quad 0 \quad \cdots]$ which clearly has a 1-norm of 1. Non-smooth regions exist for any $|\chi_i| = |\chi_j| = \|\chi\|_\infty$, $i \neq j$ which are called the *maximal equality subspaces* here. Let χ_{ns} denote a candidate point in one of these maximal equality subspaces. Define the *active coordinate vector* $\mathcal{A}(\chi_{ns}) \triangleq [\mathcal{A}_1 \quad \cdots \quad \mathcal{A}_m]^T$ as follows

$$\mathcal{A}_i(\chi) = \begin{cases} 0, & \text{if } |\chi_i| < \|\chi\|_\infty \\ \text{sign}(\chi_i), & \text{if } |\chi_i| = \|\chi\|_\infty \end{cases}$$

and the *binary permutation vector* as $\mathcal{B}(\chi_{ns}) = [\mathcal{B}_1 \quad \cdots \quad \mathcal{B}_m]^T$ where $\mathcal{B}_i \in \{0, 1\}$, $\forall \ i = 1, \ldots, m$ are the permutations of directions available for χ to increase along its maximal equality subspaces. In the neighborhood of a non-smooth point χ_{ns}, the norm may be modeled as the linear functional given by

$$\|\chi\|_\infty = \frac{(\mathcal{A}(\chi_{ns}) \circ \mathcal{B}(\chi_{ns}))^T}{\|\mathcal{A}(\chi_{ns}) \circ \mathcal{B}(\chi_{ns})\|_1} \chi.$$

Therefore, all of the possible directions of maximal ascent in an immediate neighborhood of a maximal equality subspace are then

$$\frac{d}{d\chi} (\|\chi\|_\infty) \Big|_{\chi = \chi_{ns}} \in \left\{ \frac{\mathcal{A}(\chi_{ns}) \circ \mathcal{B}(\chi_{ns})}{\|\mathcal{A}(\chi_{ns}) \circ \mathcal{B}(\chi_{ns})\|_1} \right\} \backslash \{0\}.$$

Now it is clear that the 1-norm of each of these possible gradients is equal to one. ∎

Now if we take the q-norm of (17) and subsequently apply Lemma 1 along with (16) we obtain $\|\mathbf{\Gamma}_i\|_q \leq \alpha_i \leq C$ from which we conclude (without loss of generality) that

$$\alpha_i = \|\mathbf{\Gamma}_i\|_q \leq C . \qquad \cdot (18)$$

because it may be shown that $\alpha_i > \|\mathbf{\Gamma}_i\|_q$ is always suboptimal. We may now substitute (18), (12), (13), (14) and (15) into the Lagrangian problem (11) to obtain the dual problem expressed in terms of $\{\mathbf{\Gamma}_i\}_1^\ell$ as

$$\text{Max:}_{\{\mathbf{\Gamma}_i\}_1^\ell} \quad D = -\frac{1}{2} \sum_{i,j=1}^\ell \mathbf{\Gamma}_i^T \mathbf{\Gamma}_j k(\mathbf{x}_i, \mathbf{x}_j)$$
$$+ \sum_{i=1}^\ell \mathbf{y}_i^T \mathbf{\Gamma}_i - \varepsilon \sum_{i=1}^\ell \|\mathbf{\Gamma}_i\|_q \qquad (19)$$

$$\text{S.T.:} \quad \sum_{i=1}^\ell \mathbf{\Gamma}_i = 0, \quad \|\mathbf{\Gamma}_i\|_q \leq C .$$

Here we note a similar structure to the scalar-valued case as shown in (4) and upon careful examination of (19) and (4) we observe that they are identical when $m = 1$, thus (19) generalizes the SV-SVR problem. We will now develop the KKT conditions.

C. Karush-Kuhn-Tucker (KKT) Conditions

In the SVM literature the KKT conditions state that at the optimum the product of each Lagrange multiplier and its associated constraint must vanish. For our particular problem the KKT conditions indicate that terms III through VIII of (11) must each vanish at the optimum. Let the error be defined as $\mathbf{e}_i \triangleq \mathbf{y}_i - \mathbf{W}\phi(\mathbf{x}_i) - \mathbf{b}$ and let the elements of the vectors be as follows $\gamma_i^{(*)} \triangleq \left[\gamma_{i,1}^{(*)} \quad \cdots \quad \gamma_{i,m}^{(*)} \right]^T$, $\delta_i^{(*)} \triangleq \left[\delta_{i,1}^{(*)} \quad \cdots \quad \delta_{i,m}^{(*)} \right]^T$ and $\mathbf{e}_i \triangleq \left[e_{i,1} \quad \cdots \quad e_{i,m} \right]^T$. We begin by stating without proof the rather obvious fact that $\delta_{i,j} \delta_{i,j}^* = 0$ from (10). Likewise by (11.V) and (11.VI) we have $\gamma_{i,j} \gamma_{i,j}^* = 0$. Also by the construction of (10), we may choose without loss of generality that $\delta_i - \delta_i^* \equiv \mathbf{e}_i$. Furthermore, when $\alpha_i = \|\mathbf{\Gamma}_i\|_q \neq 0$ according to (11.V), (11.VI), (11.VII) and (11.VIII) we have

$$\frac{\mathbf{\Gamma}_i}{\|\mathbf{\Gamma}_i\|_q} = \frac{\partial}{\partial \mathbf{e}_i} \left(\|\mathbf{e}_i\|_p \right) \equiv \text{sign}(\mathbf{e}_i) \left(\frac{|\mathbf{e}_i|}{\|\mathbf{e}_i\|_p} \right)^{p-1} \qquad (20)$$

which for $1 < p < \infty$ becomes $\left(\frac{|\mathbf{\Gamma}_i|}{\|\mathbf{\Gamma}_i\|_q} \right)^q = \left(\frac{|\mathbf{e}_i|}{\|\mathbf{e}_i\|_p} \right)^p$. This implies that there exists a *directional* relationship between the error $\mathbf{e}_i$ and the Lagrange multiplier $\mathbf{\Gamma}_i$ when $\|\mathbf{\Gamma}_i\|_q \neq 0$. In addition to these directional relationships, the *magnitude* of the dual variable $\mathbf{\Gamma}_i$ in vector-valued case yields information regarding the *magnitude* of the error $\mathbf{e}_i$. To explore this relationship, consider the three cases of $\alpha_i = \|\mathbf{\Gamma}_i\|_q$ with respect to its constraints at 0 and C as shown in (16).

Vanishing $\|\mathbf{\Gamma}_i\|_q$. First, consider the case where $\|\mathbf{\Gamma}_i\|_q = \alpha_i = 0$ which implies that $\|\mathbf{e}_i\|_p - \xi_i - \varepsilon \neq 0$ by (11.III) and since $\xi_i = 0$ by (11.IV) and (15), it follows that $\|\mathbf{e}_i\|_p < \varepsilon$ since $\|\mathbf{e}_i\|_p > \varepsilon$ would violate the constraint in (10). Therefore, we conclude that $\|\mathbf{\Gamma}_i\|_q = 0 \implies \|\mathbf{e}_i\|_p < \varepsilon$.

Unconstrained $\|\mathbf{\Gamma}_i\|_q$. Next, consider the case where $\|\mathbf{\Gamma}_i\|_q = \alpha_i \in (0, C)$. Since $\alpha_i \neq 0$ then $\|\mathbf{e}_i\|_p - \xi_i - \varepsilon = 0$ by (11.III). Also, since $\alpha_i \neq C$ it follows that $\eta_i \neq 0$ by (15) which in turn implies that $\xi_i = 0$ from (11.IV). Hence, we conclude that $\|\mathbf{\Gamma}_i\|_q \in (0, C) \implies \|\mathbf{e}_i\|_p = \varepsilon$.

Bounded $\|\mathbf{\Gamma}_i\|_q$. Finally, consider the case where $\|\mathbf{\Gamma}_i\|_q = \alpha_i = C$. This implies that $\eta_i = 0$ by (15), consequently $\xi_i \neq 0$ from (11.IV) and due to the constraint in (10), $\xi_i > 0$. Additionally, since $\alpha_i = C$, it follows that $\|\mathbf{e}_i\|_p - \xi_i - \varepsilon = 0$ according to (11.III) which implies $\|\mathbf{e}_i\|_p = \xi_i + \varepsilon > \varepsilon$. Hence, it is clear that $\|\mathbf{\Gamma}_i\|_q = C \implies \|\mathbf{e}_i\|_p > \varepsilon$.

D. Determining the Bias

The VV-SVM optimization problem is solved in the dual space of Lagrange multipliers, $\{\mathbf{\Gamma}_i\}_1^\ell$, leaving the bias, **b**, from the primal problem (10) yet to be determined. Just like the scalar-valued SVM, **b** is completely determined by $\{\mathbf{\Gamma}_i\}_1^\ell$ based on the KKT conditions just derived. Let the support vectors be those input vectors $\mathbf{x}_i$ for which $\|\mathbf{\Gamma}_i\|_q \neq 0$, then for each support vector which is on the margin ($\|\mathbf{\Gamma}_i\|_q \in (0, C)$) we know that the magnitude of the error is given by $\|\mathbf{e}_i\|_p = \varepsilon$ and that the direction is given by (20). Let the *biased error* be $\mathbf{F}_i \triangleq \mathbf{e}_i + \mathbf{b} = \mathbf{y}_i - \sum_{j=1}^\ell \mathbf{\Gamma}_j k(\mathbf{x}_j, \mathbf{x}_i)$ and the *signature* be $\boldsymbol{\sigma}_i \triangleq \text{sign}(\mathbf{\Gamma}_i)$ where $\text{sign}(\cdot)$ is taken element-wise and $\text{sign}(0) = 0$, then for all $i \in \mathcal{M} \triangleq \{i : \|\mathbf{\Gamma}_i\|_q \in (0, C)\}$ and $1 < p < \infty$ the KKT conditions require that

$$\mathbf{b} = \mathbf{F}_i - \varepsilon \boldsymbol{\sigma}_i \circ \left(\frac{|\mathbf{\Gamma}_i|}{\|\mathbf{\Gamma}_i\|_q} \right)^{q-1} \tag{21}$$

which allows the bias to be calculated from any element in $\mathcal{M}$. Note that this method will not work for $p = 1$ or $p \sim \infty$ because (20) does not fully convey all of the necessary direction information to properly assess the bias. In these cases, one may have to use up to m points from $\mathcal{M}$ with linearly independent signatures to determine the bias.

IV. Specific Formulations for Common Norms

The results presented thus far have been derived for the general case $1 \leq p \lesssim \infty$ which is primarily of theoretical interest. For practical computational interests, values other than $p = 1, 2, \infty$ are of less value due to their complexity. The cases of $p = 1$ and $p \sim \infty$ are appealing because they result in linear math. The case of $p = 2$ is appealing because it is Euclidian, results in a symmetry between the primal and dual spaces (since $q = 2$) and is mathematically tractable. In this section each of these three cases are studied with regard to a solution strategy.

A. 1-Norm

In this case we have $p = 1$, therefore, $q \sim \infty$ in (19) and re-introducing α_i, the dual problem becomes

$$\underset{\{\mathbf{\Gamma}_i, \alpha_i\}}{\text{Max:}} \quad D = -\frac{1}{2} \sum_{i,j=1}^\ell \mathbf{\Gamma}_i^T \mathbf{\Gamma}_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$+ \sum_{i=1}^\ell \mathbf{y}_i^T \mathbf{\Gamma}_i - \varepsilon \sum_{i=1}^\ell \alpha_i$$

$$\text{S.T.:} \quad \sum_{i=1}^\ell \mathbf{\Gamma}_i = 0, \; -\alpha_i \mathbf{1} \leq \mathbf{\Gamma}_i \leq \alpha_i \mathbf{1}, \; 0 \leq \alpha_i \leq C$$

which is quadratic in its objective and linear in its constraints. It can be solved with standard quadratic programming software. For each support vector which is on the margin ($i \in \mathcal{M}$) the KKT conditions indicate that $\|\mathbf{e}_i\|_1 = \varepsilon$. This information must be exploited to find the bias **b** because (21) cannot be used for $p = 1$.

To determine the bias, m marginal support vectors must be found which have linearly independent signatures, $\boldsymbol{\sigma}_i$. So for $i \in \mathcal{M}$, $\|\mathbf{e}_i\|_1 = \varepsilon$ may be computed as $\boldsymbol{\sigma}_i^T \mathbf{e}_i = \varepsilon$ hence it follows that

$$\boldsymbol{\sigma}_i^T \mathbf{b} = \boldsymbol{\sigma}_i^T \mathbf{F}_i - \varepsilon$$
$$\boldsymbol{\sigma}_i^T \mathbf{b} = F_i^\sigma - \varepsilon$$

where $F_i^\sigma \triangleq \boldsymbol{\sigma}_i^T \mathbf{F}_i$. Amassing all k samples in $\mathcal{M}$, a consistent system of over-determined equations is obtained as

$$\begin{bmatrix} \boldsymbol{\sigma}_1^T \\ \vdots \\ \boldsymbol{\sigma}_k^T \end{bmatrix} \mathbf{b} = \begin{bmatrix} F_1^\sigma \\ \vdots \\ F_k^\sigma \end{bmatrix} - \varepsilon \mathbf{1}$$

$$\mathcal{S} \mathbf{b} = \mathbf{F}^\sigma - \varepsilon \mathbf{1}$$

where $\mathcal{S} \triangleq \begin{bmatrix} \boldsymbol{\sigma}_1 & \cdots & \boldsymbol{\sigma}_k \end{bmatrix}^T$ and $\mathbf{F}^\sigma \triangleq \begin{bmatrix} F_1^\sigma & \cdots & F_k^\sigma \end{bmatrix}^T$ are introduced. This system is easily solved by one of two methods. Since the system of equations is consistent, m independent rows may be extracted from $\mathcal{S}$ and the equation solved directly or the entire matrix $\mathcal{S}$ may be inverted using the Moore-Penrose pseudo-inverse

$$\mathbf{b} = \mathcal{S}^+ \left(\mathbf{F}^\sigma - \varepsilon \mathbf{1} \right)$$

where $\mathcal{S}^+ \triangleq \left(\mathcal{S}^T \mathcal{S} \right)^{-1} \mathcal{S}^T$.

B. 2-Norm

In this case $p = 2$, therefore, $q = 2$. By re-introducing α_i, the dual problem becomes

$$\underset{\{\mathbf{\Gamma}_i, \alpha_i\}}{\text{Max:}} \quad D = -\frac{1}{2} \sum_{i,j=1}^\ell \mathbf{\Gamma}_i^T \mathbf{\Gamma}_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$+ \sum_{i=1}^\ell \mathbf{y}_i^T \mathbf{\Gamma}_i - \varepsilon \sum_{i=1}^\ell \alpha_i$$

$$\text{S.T.:} \quad \sum_{i=1}^\ell \mathbf{\Gamma}_i = 0, \; \mathbf{\Gamma}_i^T \mathbf{\Gamma}_i \leq \alpha_i^2, \; 0 \leq \alpha_i \leq C$$

which is quadratic in its objective but nonlinear in its constraints. It must be solved using general nonlinear programming software. Fortunately, the objective and the constraints are smooth, so gradient information is available to be used in the optimization process. For each support vector which is on the margin, the KKT conditions indicate that $\|e_i\|_2 = \varepsilon$. This information may be exploited to find the bias $\mathbf{b}$ but the use of Equation (21) is permitted. So in this case the following holds

$$\mathbf{b} = \mathbf{F}_i - \left(\frac{\mathbf{\Gamma}_i}{\|\mathbf{\Gamma}_i\|_2}\right)\varepsilon, \qquad \forall \ i \in \mathcal{M}$$

where the signature σ_i from (21) is not needed because $q - 1 = 1$ is whole and odd which preserves the sign information in $\mathbf{\Gamma}_i$.

C. ∞-Norm

In this case $p = \infty$, therefore, $q = 1$. By re-introducing γ_i and γ_i^*, the dual problem becomes

$$\begin{aligned}
\underset{\{\gamma_i, \gamma_i^*\}}{\text{Max:}} \quad D = &-\frac{1}{2} \sum_{i,j=1}^{\ell} (\gamma_i - \gamma_i^*)^T (\gamma_j - \gamma_j^*) k(\mathbf{x}_i, \mathbf{x}_j) \\
&+ \sum_{i=1}^{\ell} \mathbf{y}_i^T (\gamma_i - \gamma_i^*) - \varepsilon \sum_{i=1}^{\ell} \mathbf{1}^T (\gamma_i + \gamma_i^*)
\end{aligned}$$

$$\text{S.T.:} \quad \sum_{i=1}^{\ell} (\gamma_i - \gamma_i^*) = 0, \ \mathbf{1}^T(\gamma_i + \gamma_i^*) \le C,$$

$$\gamma_i \ge \mathbf{0}, \ \gamma_i^* \ge \mathbf{0}$$

which is quadratic in its objective and linear in its constraints. It can be solved with standard quadratic programming software. For each support vector which is on the margin, we know that $\|e_i\|_\infty = \varepsilon$. This information must be exploited to find the bias $\mathbf{b}$ because (21) cannot be used for $p \sim \infty$.

Due to the nature of the optimization problem, $\mathbf{\Gamma}_i$ will typically be sparse because one component will be more effective at increasing the objective function than the others. It is also seen from (20) that $\frac{|\Gamma_i|}{\|\Gamma_i\|_1} = \lim_{p \to \infty} \left(\frac{|e_i|}{\|e_i\|_\infty}\right)^{p-1}$ vice so for any $e_{i,j} \ne \|e_i\|_\infty$ it follows that $\Gamma_{i,j} = 0$ and visa versa. So for any element of $\Gamma_{i,j} = 0$ no conclusion may be drawn with regard to $e_{i,j}$ other than $e_{i,j} < \|e_i\|_\infty$. Then to determine the bias, at most m support vectors must be found which have linearly independent signatures. So for $i \in \mathcal{M}$, $\|e_i\|_\infty = \varepsilon$ is equivalent to $\sigma_i \circ e_i = \varepsilon |\sigma_i|$ hence it follows that

$$\sigma_i \circ \mathbf{b} = \sigma_i \circ \mathbf{F}_i - \varepsilon|\sigma_i|$$
$$\text{diag}(\sigma_i)\mathbf{b} = \mathbf{F}_i^\sigma - \varepsilon|\sigma_i|$$

where $\mathbf{F}_i^\sigma \triangleq \sigma_i \circ \mathbf{F}_i$. Amassing all k samples in $\mathcal{M}$, a consistent system of over-determined equations is obtained as

$$\begin{bmatrix} \text{diag}(\sigma_1) \\ \vdots \\ \text{diag}(\sigma_k) \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{F}_1^\sigma - \varepsilon|\sigma_1| \\ \vdots \\ \mathbf{F}_k^\sigma - \varepsilon|\sigma_k| \end{bmatrix}$$

$$\mathcal{S}\mathbf{b} = \mathbf{F}^\mathcal{S} - \varepsilon|\sigma|$$

where $\mathcal{S} \triangleq \begin{bmatrix} \text{diag}(\sigma_1) & \cdots & \text{diag}(\sigma_k) \end{bmatrix}^T$, $\mathbf{F}^\mathcal{S} \triangleq \begin{bmatrix} \mathbf{F}_1^{\sigma T} & \cdots & \mathbf{F}_k^{\sigma T} \end{bmatrix}^T$ and $\sigma \triangleq \begin{bmatrix} \sigma_1^T & \cdots & \sigma_k^T \end{bmatrix}^T$ are introduced. This system is easily solved by one of two methods. Since the system of equations is consistent, m independent (and non-zero) rows may be extracted from $\mathcal{S}$ and the equation solved directly or the entire matrix $\mathcal{S}$ may be inverted using the Moore-Penrose pseudoinverse yielding $\mathbf{b} = \mathcal{S}^+ \left(\mathbf{F}^\mathcal{S} - \varepsilon|\sigma|\right)$.

V. EXPERIMENTAL DEMONSTRATIONS

The first of two examples concerns learning a mapping $\mathbf{f} : \mathbb{R}^1 \to \mathbb{R}^2$ given by $\mathbf{y}(x) = \begin{bmatrix} e^{0.1x} \text{sinc}(x) & \cos(0.1x^2) \end{bmatrix}^T$ which is suitable for visualization. We choose 50 equally spaced samples on $x \in [0, 10]$ as the input data, a RBF kernel with $\gamma = \frac{1}{2}$, $C = 100$ and $\varepsilon = 0.1$. The results of the VV-SVR training with $p = 1$, $p = 2$ and $p \sim \infty$ are shown in Figures 1, 2 and 3 respectively. Each of these figures contain four plots illustrating the solution. For the three different norms, there were 16, 16 and 19 support vectors found respectively.

The second example is a fit of the Hwang data set [11] (which is available at the Delve database [12]) which consists of a function $\mathbf{H} : [0,1]^2 \mapsto \mathbb{R}^5$. Our intention here is to demonstrate the sparsisity of the VV-SVR approach as compared to the aggregated SV-SVR approach. In this case we use a sample size of $\ell = 2,000$. For the VV-SVR we choose $\varepsilon = 0.5$ and $p = q = 2$. To obtain a fair comparison we choose a compatible value of ε for the SV-SVM by assuring that the hyper-volume of the hyper-cube $\{\mathbf{e} : -\varepsilon\mathbf{1} \le \mathbf{e} \le \varepsilon\mathbf{1}\}$ be the same as the hyper-volume of the ball $\{\mathbf{e} : \|\mathbf{e}\|_2 \le 0.5\}$. We therefore choose $\varepsilon = 0.3485$ for the scalar-valued case. Upon performing the calculations it was found that the VV-SVR method is indeed sparser in support vectors than the aggregated SV-SVRs (which used LIBSVM [13]). Of the 2,000 training points, both methods discovered a total of 124 unique support vectors between them, 55 for the VV-SVR method and 6, 20, 28, 29 and 25 for $H_1(\mathbf{x})$ through $H_5(\mathbf{x})$ respectively for 92 unique values for the aggregated SVR method. In both cases we chose a RBF kernel with $\gamma = 8$ and $C = 100$. We observe that each SV-SVR is individually sparser than the VV-SVR, however in aggregate, they are less sparse than the VV-SVR method. The sparseness of the VV-SVR is attributable to the third term in the right hand side of (19). This term adds a cusp to the objective function which "traps" some $\mathbf{\Gamma}_i$ at $\mathbf{0}$, thus resulting in aggregate sparsity. For estimators with large dimensional input spaces, the kernel evaluation becomes significant in the computation of the estimate; it is therefore desirable to obtain the sparsest solution in terms of support vectors for efficiency of evaluation. It is in this regard that VV-SVR has an advantage over the aggregated SV-SVR approach. These sparsisity results are illustrated in Figure 4 where are shown the VV-SVR support vectors (left), the scalar-valued SVM support vectors (center) and the aggregated scalar-valued SVM support vectors (right).
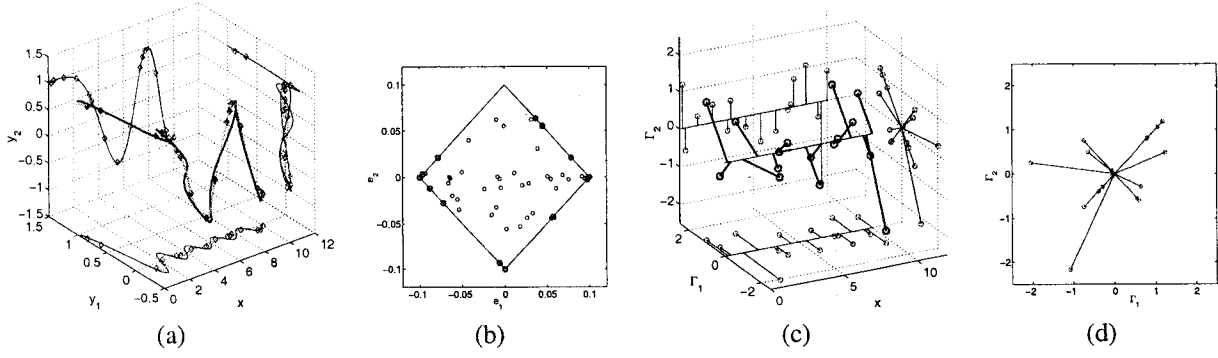
Fig. 1. 1-norm VV-SVR Approximation of $y(x)$. (a) Original function $y_i(x)$ and $\hat{y}_i(x)$. (b) Errors e_i and the ball $\|e\|_1 = \varepsilon$. (c) Lagrange multipliers Γ_i vs. x. (d) Lagrange multipliers.
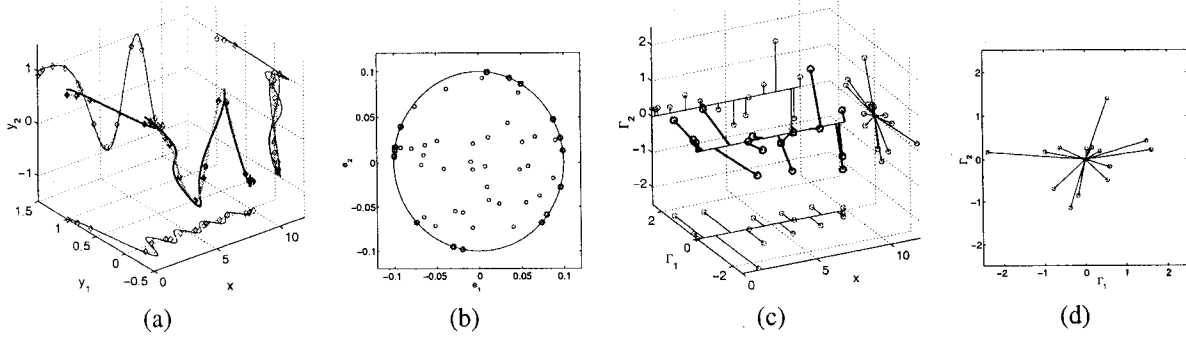

Fig. 2. 2-norm VV-SVR Approximation of $y(x)$. (a) Original function $y_i(x)$ and $\hat{y}_i(x)$. (b) Errors e_i and the sphere $\|e\|_2 = \varepsilon$. (c) Lagrange multipliers Γ_i vs. x. (d) Lagrange multipliers.
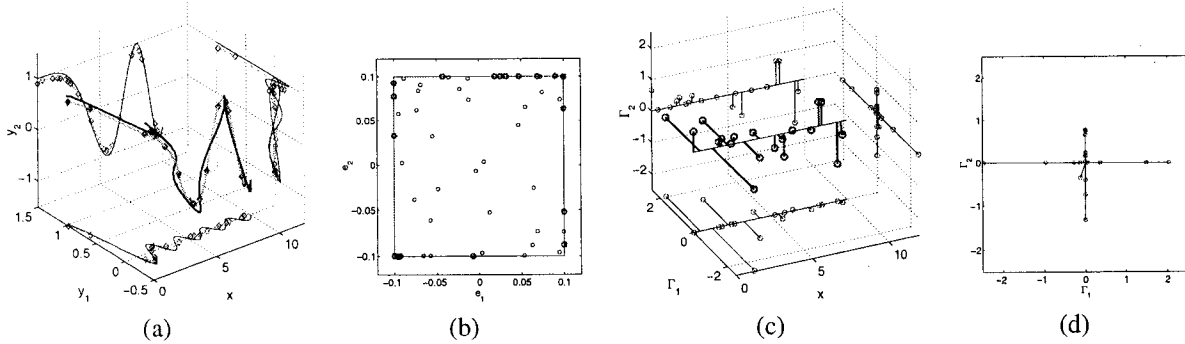

Fig. 3. ∞-norm VV-SVR Approximation of $y(x)$. (a) Original function $y_i(x)$ and $\hat{y}_i(x)$. (b) Errors e_i and the ball $\|e\|_\infty = \varepsilon$. (c) Lagrange multipliers Γ_i vs. x. (d) Lagrange multipliers.

VI. OBSERVATIONS AND CONCLUSIONS

First we observe that VV-SVR is an extension of the SV-SVR in that they are equivalent when $m = 1$. Table I shows a comparison of the two methods. Secondly we observe that the aggregated SV-SVR approach is equivalent to the VV-SVR if in (8) we let $L(e) \triangleq \||e|_\varepsilon\|_1$. Thirdly we observe that dual variables which are at bound (i.e. $\|\Gamma_i\|_q = C$) retain $m-1$ degrees of freedom. Finally, we conclude that the advantages of VV-SVR proceed from the fact that they result in sparser solutions and thus more efficient implementations.

REFERENCES

[1] C. A. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Computation*, vol. 17, no. 1, pp. 177–204, January 2005.

[2] E. Vazquez and E. Walter, "Multi-output support vector regression," in *SYSID 2003 IFAC Conference Proceedings*, 2003.

[3] C. A. Micchelli and M. Pontil, "Kernels for multi–task learning," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 921–928.

[4] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," in *16th Annual COLT Proceedings*, B. Schölkopf and M. Warmuth, Eds. Heidelberg: Springer, 2003.

[5] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM Press, April 2004, pp. 109–117.

[6] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks

TABLE I

COMPARISON OF SV-SVR TO VV-SVR.

	SV-SVR	VV-SVR
loss function	$\lvert\cdot\rvert_\varepsilon$	$\left\lvert \lVert\cdot\rVert_p \right\rvert_\varepsilon$
regularization	$\frac{1}{2}\langle\mathbf{w},\mathbf{w}\rangle$	$\frac{1}{2}\mathrm{Tr}(\mathbf{WW}^T)$
estimator (primal)	$\langle\mathbf{w},\phi(\mathbf{x})\rangle + b$	$\mathbf{W}\phi(\mathbf{x}) + b$
estimator (dual)	$\sum_{i\in I_{SV}}\beta_i k(\mathbf{x}_i,\mathbf{x}) + b$	$\sum_{i\in I_{SV}}\mathbf{\Gamma}_i k(\mathbf{x}_i,\mathbf{x}) + \mathbf{b}$
dual problem	(4)	(19)
KKT conditions	$\mathrm{sign}(\beta_i) = \mathrm{sign}(e_i)$ $\lvert\beta_i\rvert = 0 \implies \lvert e_i\rvert < \varepsilon$ $0 < \lvert\beta_i\rvert < C \implies \lvert e_i\rvert = \varepsilon$ $\lvert\beta_i\rvert = C \implies \lvert e_i\rvert > \varepsilon$	$\frac{\mathbf{\Gamma}_i}{\lVert\mathbf{\Gamma}_i\rVert_q} = \frac{\partial}{\partial e_i}\lVert\mathbf{e}_i\rVert_p$ $\lVert\mathbf{\Gamma}_i\rVert_q = 0 \implies \lVert\mathbf{e}_i\rVert_p < \varepsilon$ $0 < \lVert\mathbf{\Gamma}_i\rVert_q < C \implies \lVert\mathbf{e}_i\rVert_p = \varepsilon$ $\lVert\mathbf{\Gamma}_i\rVert_q = C \implies \lVert\mathbf{e}_i\rVert_p > \varepsilon$

with kernel methods," *Journal of Machine Learning Research*, vol. 6, pp. 615–637, April 2005.

[7] F. Pérez-Cruz, G. Camps-Valls, E. Soria-Olivas, J. J. Pérez-Ruixo, and A. R. F.-V. A. Artés-Rodríguez, "Multi-dimensional function approximation and regression estimation," in *Proc. ICANN*, Matrid, Spain, 2002.

[8] M. Sánchez-Fernández, M. de Prado-Cumplido, J. Arenas-García, and F. Pérez-Cruz, "SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2298–2307, Aug 2004.

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press, 2000.

[10] A. Smola and B. Schölkopf, "A tutorial on support vector regression," Royal Holloway College, University of London, UK, NeuroCOLT Technical Report NC-TR-98-030, 1998.

[11] J.-N. Hwang, S.-R. Lay, M. Maechler, R. D. Martin, and J. Schimert, "Regression modeling in back-propagation and projection pursuit learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 3, pp. 342–353, 1994.

[12] D. o. C. S. University of Toronto, "Data for evaluating learning in valid experiments (DELVE)," 1998, http://www.cs.toronto.edu/~delve/.

[13] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Mark Brudnak (M'04) received his B.S. in electrical engineering from Lawrence Technological University, Southfield, Michigan in 1991, his M.S. in electrical and computer engineering and his Ph.D. in systems engineering from Oakland University, Rochester, Michigan in 1996 and 2005 respectively.

Dr. Brudnak is a research engineer at the U.S. Army Tank Automotive Research Development and Engineering Center (TARDEC) which is a component of the Research Development and Engineering Command (RDECOM). His work involves the use of motion base simulators in both the durability testing of vehicle systems and the evaluation and assessment of human performance/behavior in an immersive virtual environment. His research interests involve the control and black-box modeling dynamical systems.
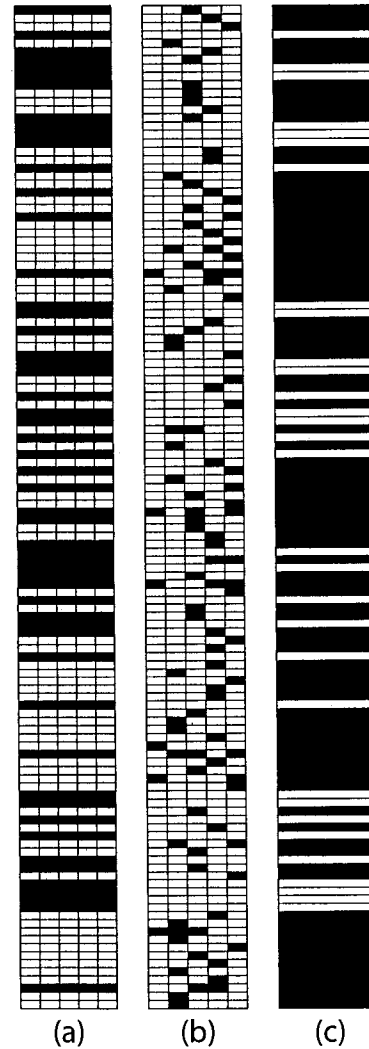
Fig. 4. Sparsisity of VV-SVR vs. SV-SVR. Shown are 124 unique support vectors found. The rows represent unique indices i and the columns indicate the outputs $H_1(\mathbf{x})$ through $H_5(\mathbf{x})$ from left to right. A black cell indicates a support vector. (a) The 55 support vectors for the VV-SVR method. (b) Plot of the support vectors found for each SV-SVR. (c) The 92 support vectors for the aggregated SV-SVR method.